Devan R. Donaldson. The Perfect Match: The Relationship between User Vocabulary and Metadata Vocabulary and Enterprise Web Information Retrieval. A Master's Paper for the M.S. in L.S. degree. April, 2008. 56 pages. Advisor: Jane Greenberg

This study examines the compatibility of user vocabulary and metadata vocabulary and then analyzes that relationship in the context of enterprise web information retrieval. An enterprise was chosen, an American Research Intensive University, and a sample from a defined set of enterprise users, undergraduates, were recruited. A quasi-experiment was conducted in which twenty subjects were asked to view ten web pages selected from the enterprise's website and write a description of each page. Terms written by subjects were compared with metadata terms assigned to the web pages by professionals. Instances of exact, partial and no match between the two vocabularies were recorded. Searches were conducted using subjects' terms via the enterprise's web search engine. The position of selected web pages in search engine results lists were recorded upon completion of each search. Results suggest a statistically significant relationship between the vocabulary matches and enterprise web information retrieval.

Headings:

Metadata

Information Organization

Information Retrieval

THE PERFECT MATCH: THE RELATIONSHIP BETWEEN USER VOCABULARY AND METADATA VOCABULARY AND ENTERPRISE WEB INFORMATION RETRIEVAL

by
Devan R. Donaldson

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

April 2008

Approved by

_____

Jane Greenberg

TABLE OF CONTENTS

1. Introduction 3

1.1 Broad Purpose and Key Terms 4

1.2 Uniqueness of Study and its Importance 5

1.3 Why Solicit and Analyze User Vocabulary? 5

1.4 Why an Enterprise? 6

1.5 The Research Question and Hypotheses 7

1.6 Specifics of What was Done and What was Found 8

2. Literature Review 10

2.1 The Debate 10

2.2 Is Metadata Effective for Web Information Retrieval? 12

2.3 Certain Metadata is More Effective for Web Information Retrieval than Others .. 15

2.4 Metadata is Not Effective for Web Information Retrieval 16

2.5 Metadata in Enterprises Presents a Different Story 17

2.6 What Makes Metadata Effective? 17

2.7 Conclusion 18

3. Study Design and Method 20

4. Findings/Results 25

4.1 Statistical Analysis 25

4.2 User Vocabulary and Metadata Vocabulary Compared: Issues Resulting from Differences in Spacing, Spelling, and Other Types of Variations 31

4.2.1 Web pages 3 and 6 31

4.2.2 Concerning Parts of Words and Abbreviations and Visibility in Search Engine Results Lists 32

4.2.3 Addition of Terms at Times Made EWIR less Effective 33

4.2.4 Concerning Use of Alternate/Generic Terms 34

4.2.5 Regarding Variation in Terms of Spacing 35

4.2.6 Conclusion – Should All of this be Occurring? 35

5. Summary and Conclusions 37

5.1 Summary 37

5.2 Conclusions 38

Works Cited 40

Appendices 1-11 42

1. Introduction

Millions upon millions of people are using the World Wide Web to search for information every day. The search engine is one of the most popular tools designed to help facilitate the process of finding information. Search engines are defined as "[w]eb sites or software that search the Internet for documents that contain a key word, phrase, or subject that is specified by the user to the search engine" ("search engine" 1). Metadata, most commonly defined as data about data, is thought to be created for the purpose of resource discovery. In the context of web search, metadata is thought to help search engines match users with web pages based on the terms supplied by users. Specifically, metadata, either human-created or automatically created, about web pages is supposed to match with user terms. That way, when users provide terms to search engines, search engines will be able to match users with desired content. Thus, match between user terms and metadata terms is needed for users, and the information they seek, to unite. Without studying to make sure that: 1) user terms and metadata terms are matching, and 2) match between both vocabularies results in users finding the information they are in search of, it could be that the work of many metadata creators is in vain and many web searchers are sitting at home or work frustrated by futile attempts to find the information they need. This study does not presuppose that what is "supposed to happen" regarding user terms, metadata, and search engines is actually occurring.

Rather, this study was designed to investigate what is actually happening among user terms, metadata, and search engines.

In this chapter, the purpose of the study reported in this paper will be discussed along with provision of definitions of key terms as they are used within this paper. Following this, information concerning what makes this study unique and why it might be considered important is provided. Afterwards, the question of why user vocabulary might be solicited and analyzed is addressed. Why choosing an enterprise for such a study is also mentioned. Finally, presentation of the research question, hypotheses, and an overview of the study design, method, and findings are provided.

1.1 Broad Purpose and Key Terms

The purpose of this study is to better understand the relationship between user vocabulary and metadata vocabulary as well as to understand what impact, if any, that relationship has on enterprise web information retrieval. User vocabulary refers to the terms that users provide, which they think will retrieve web pages of a desired kind. Metadata vocabulary refers to the content supplied by creators of enterprise web pages within <title></title> tags as well as content supplied within the meta tags <meta name="description" content=""""""> and <meta name="keywords" content="""""">. Note that specific information concerning who created metadata for the web pages was not gathered because this was not the focus of the study. Thus, it was assumed that the metadata assigned to web pages were created by professionals—enterprise staff members with at least some level of training and experience in creating metadata for web pages.

The term enterprise refers to an entity, which provides web content (e.g, a college, university, and/or business). The phrase "enterprise web information retrieval" (EWIR) refers to the process by which web pages, designed by enterprise web page creators, are displayed via the enterprise's search engine results lists for the purpose of information retrieval.

1.2 Uniqueness of Study and its Importance

The compatibility of user vocabulary and metadata vocabulary—defined as the extent to which user and metadata vocabulary match—has never been empirically examined at the enterprise chosen for this study—until now. Also, the weight given to metadata in the enterprise's search engine algorithm is unknown to the public. Thus, only through conducting research can an understanding of the relationship between user and metadata vocabulary match/mismatch and the effectiveness of EWIR be found, if such a relationship exists at all.

1.3 Why Solicit and Analyze User Vocabulary?

A substantial body of research and scholarship exists, which addresses the issue of vocabulary match/mismatch with respect to users and information retrieval systems, such that "the keywords that are assigned by indexers are often at odds with those tried by searchers" (Furnas et. al. 965). I consider it important to place this match/mismatch issue within the context of EWIR in order to provide a context for the evaluation of match/mismatch. After all, understanding this issue in an enterprise context does not

preclude the fact that users are creating queries using terms with the expectation of finding the web pages they need.

Some researchers, such as Hawking and Zobel in "Does topic metadata help with web search?", chose to develop queries from site-maps over acquiring queries from users when trying to examine the effectiveness of metadata for EWIR (617). Hawking and Zobel felt that "the match in terminology between the site-map and the pages it indexes [was] likely to be better than the match between queries and pages sought by a user" (620). In contrast, I consider emphasis on the user as central to understanding the effectiveness of metadata for EWIR. Thus, in this study, query terms were taken from actual users. The hope was that users' terms would match metadata terms. Furthermore, it was hoped that, when user and metadata vocabulary matched, enterprise web pages would be retrieved highly visibly.

1.4 Why an Enterprise?

In this study, one enterprise was chosen, an American Research Intensive University, whose name will remain anonymous in this paper. This study was carried out using web pages from the study university's website with subjects who were current undergraduate members of that enterprise because: 1) metadata created in an enterprise environment is thought to be much more trustworthy than metadata created generally on the World Wide Web (Brooks 11; Dawson & Hamilton 310), and 2) enterprise users are a much more narrowly defined group of users than those of the World Wide Web. Most

importantly, the enterprise web pages were created specifically to meet the information needs of enterprise users.

1.5 The Research Question and Hypotheses

This study was undertaken to address the following research question: *Is the extent to which user vocabulary and metadata vocabulary match at all associated with EWIR?* To address this research question, the extent to which user and metadata vocabulary match was measured on three levels: exact, partial, and no match. EWIR was measured as page rank of web pages in enterprise search engine results lists, with 1 being the most favorable position in a search engine results list and 20 being the least favorable position in an enterprise search engine results list. Based upon the research question, the null hypothesis and other hypotheses were tested. They are:

> *Null Hypothesis*
>
> *H_0 = There is no association between the extent to which user and metadata vocabulary match and EWIR.*
>
> *Hypothesis 1*
>
> *H_1 = There is an association between the extent to which user and metadata vocabulary match and EWIR such that the more user and metadata vocabulary match, the more effective the EWIR.*
>
> *Hypothesis 2*
>
> *H_2 = There is an association between the extent to which user and metadata vocabulary match and EWIR such that the less often user and metadata vocabulary match, the less effective the EWIR.*

Thus, consider if, no matter whether user and metadata vocabulary matched exactly, partially, or not at all, EWIR was the same. This would mean there was no association

between the extent to which user and metadata vocabulary match and EWIR. This would also mean that the null hypothesis would be validated, and hypotheses 1 and 2 would be rejected. If, however, exact match coincided with web pages at the page rank of 1 in enterprise search engine results lists, which is defined as more effective EWIR, this would suggest that there is an association between the extent to which user and metadata vocabulary match and EWIR. Therefore, the null hypothesis would be rejected, and hypothesis 1 would be validated. If no match between user and metadata vocabulary coincided with web pages found at the page rank of 20, the least favorable EWIR position, this would reject the null hypothesis and validate hypothesis 2. It was hoped that the null hypothesis would be rejected and hypotheses 1 and 2 would be validated.

If, as a whole, instances of partial match resulted in more effective EWIR than in instances of exact match, or if, as a whole, instances of partial match resulted in worse EWIR than in instances of no match, this might not have an effect on the validation or rejection of the hypotheses, but would suggest that other factors, beyond the scope of this study, were at play. Thus, a hypothesis devoted to partial match was not constructed.

1.6 Specifics of What was Done and What was Found

A study was designed in order address the research question by testing the hypotheses. To this end, a sample from a defined set of enterprise users, undergraduates, was recruited. A quasi-experiment was conducted in which subjects were asked to: 1) view ten web pages, selected from the enterprise's website, and 2) write descriptions for each web page. After the quasi-experiment was completed by all subjects, the researcher

compared terms written by subjects with metadata terms. Instances of exact, partial and no match were recorded. The researcher used the subjects' terms to conduct searches using the enterprise's web search engine. (Note that the researcher, rather than the subjects, completed all searches so that subjects could focus exclusively on writing terms they would use to search for certain web pages during their time engaged in the quasi-experiment.) The position of web pages in search engine results lists was recorded upon completion of each search. Results (details follow in a later section) suggested that there is a statistically significant relationship between the extent to which user and metadata vocabulary match and EWIR. While the correlation found was strong and negative, it was only found to be "perfect" with one set of data. Thus, the researcher conducted analysis of "outliers" to further analyze the circumstances governing cases which seemed to buck the trend of the data. In these cases, user terms and phrases varied from metadata vocabulary in spacing, spelling, singular and plural form, etc., and, at times, this adversely affected the visibility of desired web pages in search engine results lists.

The chapters that follow include: a literature review, information concerning the study design and method, findings and results, and a summary and conclusions.

2. Literature Review

This chapter consists of a literature review which explores the debate among scholars regarding the effectiveness of metadata for web information retrieval (WIR). Enterprises are introduced as a counter to the circumstances under which criticism of metadata and its role for WIR have been couched. Afterwards, specific examples of empirical research regarding metadata and its effectiveness for WIR are presented, compared, and contrasted. Empirical research devoted to assessing which type(s) of metadata are most effective for WIR is also considered. Why metadata might be seen as not particularly helpful for WIR is revisited as well as the topic of metadata in enterprises. Finally, suggestions of researchers regarding what makes metadata effective for WIR are presented and concluding remarks are provided.

2.1 The Debate

A number of researchers have found that metadata enhances WIR (Zhang and Dimitroff 2004, 318). Some have even gone as far as to suggest which metadata or combinations of metadata are more effective for WIR (Brackbill and Turner 267). Other researchers have suggested that metadata is not effective for WIR, suggesting that, when web pages with and without metadata are compared, there is little difference in the visibility of those web pages in search engine results lists (Henshaw and Valauskas 89).

On the other hand, some researchers have run similar studies with statistical findings that suggest the opposite (Zhang and Dimitroff 2004, 312-316).

Still other researchers have suggested that metadata is not effective for WIR because some metadata creators misuse metadata fields and meta tags in such a way that more favorable search engine results visibility trumps accuracy of the metadata used to describe web pages (Lynch 13). This has caused many search engine providers to mistrust metadata creators, and, as a result, some search engine providers refuse to weigh metadata in their search algorithms (de Groat 21). Given all of the information provided above, it may be that metadata is effective for WIR if: 1) metadata is accurate, and 2) search engine providers weigh metadata in their search algorithms.

Accuracy of metadata is thought to be much less of a problem in the case of enterprises (Brooks 11; Hawking and Zobel 615). Enterprises, which in this paper are defined as educational institutions and government organizations and/or agencies, are comprised of a group of known metadata creators and known users who are accountable to themselves and the enterprise of which they are a part in ways that metadata creators and users of the World Wide Web are not. Therefore, the metadata associated with enterprise web pages may be more trustworthy.

Whether search engine providers weigh metadata in search algorithms depends upon the design of each search engine. No matter how good the metadata, if it is ignored in search engine algorithms, the metadata may not help to facilitate or improve WIR. If search engine providers are weighing metadata in their algorithms, the search engine

providers typically do not share any specific details about how the metadata is weighed. This is because some metadata creators would use this information to manipulate their metadata for increased visibility of web pages—regardless of whether or not the metadata is accurate in its description of web pages (Brooks 10). Creating metadata purely for search engine optimization of web pages "by any means necessary" sacrifices trust and integrity.

If metadata is created in a trustworthy environment, and metadata is used by search algorithms within a trustworthy environment, what are additional characteristics of effective metadata? Researchers suggest that metadata should reflect the terminology and interpretation of information from the user's perspective to the greatest extent possible (Brasethvik 385; Hawking and Zobel 625).

2.2 Is Metadata Effective for Web Information Retrieval?

A number of researchers have sought to examine the effectiveness of metadata for WIR by conducting searches using web pages or web resources *with* metadata and by conducting searches using web pages or web resources *without* metadata for the purpose of comparison.

In "Metadata as a catalyst: experiments with metadata and search engines in the Internet journal, First Monday," Henshaw and Valauskas present a study in which they conducted several searches using keywords extracted from the title and text of papers selected from 30 issues of *First Monday*, an internet-only journal. Search engines used in this process included AltaVista, Excite, Google, Hotbot, Infoseek, Lycos, and

Northernlight. Search results were recorded. Soon after, metadata, from the Dublin Core Metadata Initiative, were added to the same web resources, searches were repeated and results were recorded. Henshaw and Valauskas found that while some search engines demonstrated sensitivity to meta tags, there was no clear evidence that meta tags greatly enhanced the ranking of selected papers from *First Monday* (Henshaw and Valauskas 89). Specifically, they found that, with the addition of metadata, the papers were: 1) not likely to have experienced any change in ranking, and 2) more likely to have experienced loss rather than gain in ranking. However, there is a significant caveat in Henshaw and Valauskas's findings: "depending on the search engine's own timetable for examining and indexing the content of *First Monday*, not all metatagged articles were captured, hence leading to some discrepancies in search results" (90). This circumstance definitely causes problems when trying to interpret their data because there is no way to know which search engines had actually indexed the metadata that was added to the papers, and which ones had not, by the time Henshaw and Valauskas ran their second set of searches. It may very well be the case that all or none of the metadata that was added to the papers was indexed by the search engine by the time Henshaw and Valauskas ran their second set of searches.

In "Internet search engines' response to metadata Dublin Core implementation," Zhang and Dimitroff present a study in which they looked at the performance of search engines with regard to web pages with and without metadata. They: 1) created a set of web pages without metadata, 2) submitted those web pages for indexing to seven search engines (AllTheWeb, EntireWeb, Google, Lycos, AltaVista, Yahoo, and Infospace/Fast),

3) searched for the web pages, 4) recorded page rank for each web page, 5) added metadata to each web page, 6) searched for each web page with metadata added, and 7) recorded the page rank of each web page. Their null hypothesis stated that "there [would be] no significant difference for the search engines with respect to search engine visibility performance of web pages before and after the metadata Dublin Core elements [were] implemented in [the] web pages" (Zhang and Dimitroff 2004, 311). The researchers operationalized page rank on search engine results lists as a measure of effectiveness of WIR; essentially, the closer the page rank to number 1 the more effective the WIR. In their study, the independent variables were the identified search engines and the dependent variable was web page visibility in search engine results lists (Zhang and Dimitroff 2004, 312). Statistical analysis was employed to test the null hypothesis against the data for each search engine. In the case of almost every search engine, Levene's F was statistically significant. This meant that the means of the groups, web pages with metadata and those without, were different in a statistically significant way (Zhang and Dimitroff 2004, 312-316). Also, the means for web pages with metadata elements were lower than the means for web pages without metadata elements. This showed that web pages with metadata elements achieved better visibility performance than those without (Zhang and Dimitroff 2004, 312-316).

Henshaw and Valauskas and Zhang and Dimitroff ran similar studies but with different findings and conclusions. What appropriate conclusion(s) should be drawn from the fact that both studies differ in terms of their results? Furthermore, based on the

difference in results, in what way(s) should one regard metadata and its effectiveness for WIR? This sort of discrepancy provides a basic justification for this study.

2.3 Certain Metadata is More Effective for Web Information Retrieval than Others

Some researchers have conducted studies to address what kinds of metadata might be more effective for WIR than other types. In "Rising to the top: evaluating the use of the html meta-tag to improve retrieval of world wide web documents through internet search engines," Turner and Brackbill recount a study in which they designed twenty web pages in five main groups of the subject agriculture. Four pages were dedicated to each group: the first without meta tags, the second with the keywords meta tag, the third with the description meta tag, and the fourth with keywords and description meta tags. All of the web pages were submitted to AltaVista and Infoseek. Afterwards, both search engines were searched using keywords extracted from the web pages. Turner and Brackbill found that using the keywords meta tag, either with or without the description meta tag, improved retrieval rank over using only the description meta tag (267). This would suggest that keywords metadata is more effective for WIR than description metadata. This would also suggest that having more metadata, for example, keywords and description metadata, does not necessarily increase visibility in search engine results lists. Similarly to Turner and Brackbill, in "The impact of metadata implementation on web page visibility in search engine results (part II)" Zhang and Dimitroff discuss a study in which they used web pages with different numbers of metadata element combinations and found that web pages with keywords appearing in the metadata title field, metadata subject field and metadata description field achieved better visibility performance than

other possible combinations (2005, 704). The results of the aforementioned studies suggest that it is not enough to simply include metadata, but that some metadata are more effective for WIR than others.

2.4 Metadata is Not Effective for Web Information Retrieval

Other researchers have characterized metadata, regardless of type, as ineffective for WIR. This is not because good metadata has been proven to fail, but instead, because some metadata creators have sacrificed the integrity of metadata in order to make their web pages more visible. As Clifford Lynch points out, "metadata may be carefully constructed by any number of parties to manipulate the behavior of retrieval systems that use it, rather than simply describing the documents or other digital objects it may be associated with" (Lynch 13). Consequently, some search engines do not weigh meta tags such as <meta name="keywords"> and <meta name="description"> at all in search algorithms designed to populate search engine results lists because often these meta tags are misused (de Groat 21). This is unfortunate because there are metadata creators who do use the meta tags appropriately, and their efforts to improve retrieval of web pages by creating metadata are to no avail because of those who have misused meta tags. There are no metadata police to make sure that metadata creators are using meta tags correctly in the World Wide Web. Thus, in the World Wide Web, there is a lack of accountability on the part of metadata creators for the information they provide. As a result, many search engine providers neither trust metadata creators nor their metadata.

2.5 Metadata in Enterprises Presents a Different Story

In enterprises, metadata is thought to be effective for WIR. Enterprises are considered to be reputable, which has caused search engine providers to be much more trusting of metadata from ".gov" or ".edu" websites (Dawson and Hamilton 310). Furthermore, as Terrance Brooks points out, "enterprises are known to be driven by social groups that reach agreements on information structure and topical metadata as opposed to the more arbitrary and self-motivated decisions behind the creation of metadata on the World Wide Web" (11). Thus, because the integrity of the metadata is thought to reflect the integrity of the enterprise, metadata created by enterprise metadata creators is thought to be helpful for EWIR.

2.6 What Makes Metadata Effective?

More generally, researchers have made suggestions regarding what makes metadata effective and have provided essential characteristics of effective metadata. Zhang and Dimitroff, for instance, have suggested that, "[s]uccessful use of metadata to communicate meaning of information relies on users' understanding or awareness of other's interpretation of the domain and how this interpretation is reflected in the metadata statement" (Zhang and Dimitroff 2005, 693). From this it can be deduced that metadata creators should make it their business to: 1) understand how users understand and interpret given information, and 2) reflect their understanding of how users understand and interpret given information in the metadata they create. These aims cannot be accomplished unless metadata creators study users. Thus, metadata creators

should attempt to gather information from users to make sure metadata creators and users "are on the same page" in terms of their description of the same information. These efforts, on the part of metadata creators, are needed in order to assess whether or not metadata is truly effective for the purpose of WIR. Perhaps, when metadata creators study users, they will find that the terms metadata creators employ to describe web pages are exact or quite close to the terms users would employ. This would be a good thing, but without checking to see if this is actually the case, one can never know for sure.

Hawking and Zobel point out that metadata, "should add something to the data that cannot be deduced from the visible text: otherwise, users will not understand why a particular page has been retrieved, as the metadata is not displayed" (Hawking and Zobel 625). Thus, based upon suggestions from the aforementioned researchers, metadata should embody the language and interpretation of those who presumably would seek web pages—the users.

2.7 Conclusion

The literature reviewed in this chapter could all be used to suggest that more research should be undertaken to understand whether users are retrieving the web pages they would expect to find by using the terms that they consider appropriately describe the web pages they are looking for. Based on the researcher's review of literature concerning metadata and its effectiveness for WIR, he developed a study in the context of an enterprise in which: 1) user vocabulary and metadata vocabulary were compared, 2) actual user query terms and phrases were tested, and 3) the page rank of web pages for

which query terms and phrases were generated were recorded. All of this information was analyzed as a means of trying to better understand what role metadata actually plays in WIR.

3. Study Design and Method

A quasi-experiment was employed in this study. Quasi-experiments resemble controlled experiments but lack key elements such as pre- and posttesting and/or control groups (Trochim 1-6; Babbie 349). Examples of specific quasi-experiments include: 1) time-series design "which involves measurements of a given action made over time" (Babbie 349), 2) nonequivalent control group experiments in which "a control group that is similar to the experimental group [is used] but is not created by random assignment of subjects" (Babbie 350), and 3) multiple time-series design in which one set of data, that was collected over time, is used for the purpose of comparison (Babbie 352). Although the quasi-experiment that was used in this study does not possess all of the qualities outlined in the types of quasi-experiments mentioned above, it does share enough characteristics among those types of quasi-experiments to be considered as such. The quasi-experiment employed in this study neither had a control group nor pre- or posttesting, but did have a set of procedures which all subjects were expected to follow.

In this quasi-experiment, twenty undergraduates were asked to view ten selected web pages and write descriptions of each web page e.g. record the terms they would use if they were searching for these web pages via the enterprise's search engine. The ten selected web pages: 1) were taken from the same enterprise, 2) had metadata associated with them that could be viewed by way of viewing the source code of each web page, and 3) had metadata using the <title></title> tag as well as the meta tags <meta

name="description" content=""> and <meta name="keywords" content="">. In addition to the three aforementioned criteria, half of the web pages selected for this quasi-experiment were chosen from an enterprise web page designed specifically for undergraduates. Note also that these web pages were provided at the same level of scope. In other words, each web page listed was hyperlinked using the same font and colors and featured only one time in the list. The other five web pages were chosen because, in addition to having satisfied the criteria outlined above in items 1 through 3, the web pages could be argued to be relevant in at least some capacity to undergraduates. These web pages included: information about the study university's school of business, a school which does offer undergraduate degrees, two web pages from the study university's school of public health, and one web page about the study university's program in environmental sciences and engineering—which does offer a bachelor's degree. Because finding ten web pages that fit selection criteria 1 through 3 and were also related, in at least some capacity, to undergraduate life and/or services was not possible (based on the researcher's survey of the university website), one web page was chosen for the quasi-experiment which contained information concerning the study university's program in epidemiology—a program which does not offer a bachelor's degree. The epidemiology web page was used to populate the sample of web pages only because it satisfied selection criteria 1 through 3. However, it is certainly possible that the web page could be relevant to some undergraduates, despite the fact that the study university does not offer undergraduate degrees in epidemiology.

The sample of subjects was populated based upon: 1) subjects' responses to a recruitment email sent out to all of the study university's undergraduates, and 2) the researcher's recruitment of subjects in person by asking undergraduates, who were present in the study university's student union, to participate. Subjects who responded to the researcher's email expressing interest in the quasi-experiment were contacted to confirm a time to meet one-on-one to complete the quasi-experiment. Upon arriving at the main lobby of the student union, each subject was handed an informed consent fact sheet. Upon verbal agreement to participate in the study, each subject was asked to sit down in front of the researcher's personal Lenovo IBM ThinkPad X41 Tablet laptop computer, which the researcher had set up on the far end of the study university's student union lobby. The researcher handed each subject a Written Query Sheet (see Appendix 11) which had the following directions written at the top:

1) *Please type in the url of Web page 1, found on the Written Query Sheet, into the address bar of the Microsoft Internet Explorer 7.0 web browser, which is available from the computer at the computer station.*
2) *Press enter to view Web page 1.*
3) *Write the words and/or statements you would use to describe Web page 1 in the space provided on the Written Query Sheet.*
4) *Repeat steps 1 through 3 for Web pages 2 through 10.*

Please spend no more than two minutes per web page to view and write descriptions and please print your responses clearly and legibly. When you are finished, please return the Written Query Sheet to the Principal Investigator (PI).

After each subject finished, the researcher collected the Written Query Sheet and reviewed the subject's responses to make sure the responses were legible. If the researcher could not understand what a subject had written, the researcher asked the subject for clarification before that subject's study session ended. If the responses were

in fact legible, the quasi-experiment was considered complete. Afterwards, the researcher took the subject to the university's student stores or coffee shop, the choice was up to the subject, where he/she was allowed to choose either a snack or drink of his/her choice—not to exceed $5.00.

After the researcher collected completed Written Query Sheets from twenty undergraduates, the researcher used the terms (exactly as they were provided by subjects—maintaining the spelling, punctuation, spacing, capitalization, etc.) as queries and searched using the enterprise's search engine—a total of 200 searches. (Note that the researcher was unable to find out specifically how metadata was used in the enterprise's search engine algorithm, but was able to verify from sources at the enterprise's information technology services web systems department that: 1) metadata was weighed in the enterprise's search algorithm, and 2) no specific metadata terms were weighed more heavily than others.) After each search, the researcher recorded the page rank of the page for which the description was created. The researcher did not look beyond the twentieth search result. Thus, if the web page was not found in search results 1 through 20, this was designated by the number 20. Also, regardless of how closely related certain retrieved web pages were, only results for the ten selected web pages were recorded.

Afterwards, the researcher compared the terms provided by users for web pages with the metadata of each of the web pages. The metadata was available by viewing the source code of each web page. If the terms users provided matched metadata exactly, this was recorded as 1. If terms users provided matched metadata partially, this was indicated with a 0. If terms users provided did not match metadata terms, this was

indicated with a -1. The researcher used all data concerning page rank and match to calculate descriptive statistics and various correlation information.

4. Findings/Results

In this chapter, findings and results will be discussed, including: 1) statistical analysis of the data, and 2) discussion of outlier cases—which in this study will be taken to mean cases which seemed to buck the trend of the data. These cases were the result of variation between user and metadata vocabulary with respect to spacing, spelling, abbreviation, and the addition of terms.

4.1 Statistical Analysis

The relationship between user vocabulary and metadata vocabulary was defined in terms of match (*Match*), which served as the independent variable. There were three levels of the independent variable: 1, which represented exact match between user and metadata vocabulary; 0, which represented partial match between user and metadata vocabulary; and -1, which represented no match between user and metadata vocabulary.

The findings suggest that when enterprise users are looking at the same web pages created by enterprise metadata creators, more often than not, enterprise users come up with the same terms to describe web pages that enterprise metadata creators use. As Table 1 shows, there were 103 cases of exact match out of 200 searches, 51%, and 67 cases of partial match out of 200 searches, 34%. Thus, user and metadata vocabulary matched at least partially 85% of the time.

Table 1. Summary of *Match*

Selected web pages	Instances of exact match (*Match* = 1)	Instances of partial match (*Match* = 0)	Instances of no match (*Match* = -1)
1	11	8	1
2	13	6	1
3	8	5	7
4	13	5	2
5	16	4	0
6	12	6	2
7	5	9	6
8	5	9	6
9	10	6	4
10	10	9	1
TOTAL*	103	67	30
% of all searches	51%	34%	15%

*Total out of all 200 searches.

The dependent variable, EWIR, was operationalized as page rank (*Pagerank*) in enterprise search engine results lists. Essentially, after conducting a search, the closer the page rank of a given web page to number 1 in a search engine results list, the more effective the EWIR. Conversely, after conducting a search, the farther away the page rank of a given web page was from number 1 in a search engine results list, the less effective the EWIR. There were twenty levels of the dependent variable (1-20): 1, representing page rank of a web page in a search engine results list in the most favorable position possible, and 20, representing page rank of a web page in a search engine results list in the least favorable position possible.

To begin to address the primary research question, the null hypothesis, the extent to which user vocabulary and metadata vocabulary match is not at all associated with

EWIR, had to be tested against the data. To this end, bivariate correlation analysis was run to find the Pearson Correlation Coefficient (*r*) (see Table 2). Based on the data collected for eight out of ten of the selected web pages, the correlation between *Match* and *Pagerank* was found to be statistically significant. For web page 1, the correlation

Table 2. Pearson Product Correlations

Selected web pages	Pearson Correlation Coefficient (*r*)	Statistical Significance (*p*<.001)
1	*r*(18)* = -.853	0.000
2	*r*(18)* = -.892	0.000
3	*r*(18)* = -.684	0.001
4	*r*(18)* = -.793	0.000
5	*r*(18)* = -1.000	0.000
6	*r*(18)* = -.648	0.002
7	*r*(18)* = -.838	0.000
8	*r*(18)* = -.779	0.000
9	*r*(18)* = -.896	0.000
10	*r*(18)* = -.764	0.000

*The number in parentheses represents the degrees of freedom associated with the significance test, which is equal to the number of cases minus 2 (or N-2). Since each subject provided descriptions for each web page and there were twenty subjects, the number in parentheses is the same for data concerning each web page.

between *Match* and *Pagerank* was significant, $r(18) = -.853$, $p<.001$. The number in parentheses represents the degrees of freedom associated with the significance test, which is equal to the number of cases minus 2 (or N-2). Since each subject provided descriptions for each web page and there were twenty subjects, the number in parentheses is the same for data concerning each web page. For web page 2, the correlation between

Match and *Pagerank* was significant, $r(18) = -.892$, $p<.001$. For web page 4, the correlation between *Match* and *Pagerank* was significant, $r(18) = -.793$, $p<.001$. For web page 5, the correlation between *Match* and *Pagerank* was significant, $r(18) = -1.000$, $p<.001$. For web page 7, the correlation between *Match* and *Pagerank* was significant, $r(18) = -.838$, $p<.001$. For web page 8, the correlation between *Match* and *Pagerank* was significant, $r(18) = -.779$, $p<.001$. For web page 9, the correlation between *Match* and *Pagerank* was significant, $r(18) = -.896$, $p<.001$. For web page 10, the correlation between *Match* and *Pagerank* was significant, $r(18) = -.764$, $p<.001$. In each case, *r* was negative and near -1, which would suggest that low scores on *Match* tend to be associated with high scores on *Pagerank*, and high scores on *Match* tend to be associated with low scores on *Pagerank*. Based on the levels provided for the independent variable and dependent variable, this would mean that the closer *Match* is to -1 (no match between user vocabulary and metadata vocabulary), the higher the *Pagerank* (the closer the page rank is to rank 20, which is least favorable). Conversely, the closer *Match* is to 1 (exact match between user vocabulary and metadata vocabulary), the lower the *Pagerank* (the closer the page rank is to rank 1, which is most favorable). Based on the results, the null hypothesis, the extent to which user vocabulary and metadata vocabulary match is not at all associated with EWIR, was rejected. Specifically, the more user and metadata vocabulary matched, the more effective the EWIR. Thus, hypothesis 1 was validated. Also, the less often user and metadata vocabulary matched, the less effective the EWIR. Thus, hypothesis 2 was validated. All of this is exactly what one would hope, assuming that the metadata was created to aid EWIR.

For web pages 3 and 6, the correlation between *Match* and *Pagerank* was not found to be statistically significant. For web page 3, the correlation between *Match* and *Pagerank* was not statistically significant, $r(18) = -.684$, $p=.001$. For web page 6, the correlation between *Match* and *Pagerank* was not statistically significant, $r(18) = -.648$, $p=.002$. The weaker correlation observed with respect to web page 3 can be explained by two cases in which exact match, 1, was accompanied by page rank of 20, the least favorable search engine results list position. In both cases, users provided the abbreviation "ITS" as a term that they thought would most effectively retrieve the study university's information technology services web page, which was one of the ten selected web pages for the quasi-experiment. The abbreviation did match as a metadata vocabulary term exactly, as was verified by viewing the source code of the information technology services web page. Surprisingly, the information technology services web page was not visible in results 1 through 20 after searching using the term provided by both users. Although in each case the web page which was first on the search engine results list was actually an "About ITS" web page that was part of the information technology services website, this was not taken into account by the researcher, because, in this study, only the ten selected web pages were considered. In the case of data and results for web page 6, there were two cases in which no match between user and metadata vocabulary, -1, resulted in pages retrieved at the page rank of 10. In one case, the user provided the term "fraternities" as a term the user thought would retrieve the study university's Greek life web page. In another case, the user provided the term "sororities" as a term the user thought would retrieve the Greek life web page. Although

neither “fraternities” nor “sororities” matched metadata vocabulary, both terms, in their singular form, were included as metadata vocabulary. Perhaps because the singular and plural forms of the same terms are, by their nature, closely related, the web page for which the terms were provided showed up in search engine results lists fairly visibly—albeit more modestly than otherwise. Nevertheless, the Pearson Correlation Coefficients for web pages 3 and 6 are moderately negative and could still be used to support the suggestion that the extent to which user vocabulary and metadata vocabulary match *is* associated with EWIR—in the same way that was suggested based upon statistically significant results from the data that was collected for the other web pages selected for this quasi-experiment.

Assuming that the independent variable, *Match*, is thought of as the predictor and the dependent variable, *Pagerank*, as the criterion, the Pearson Correlation Coefficients were squared to determine the strength of the relationship between the variables based on data collected for all ten selected web pages: for data collected concerning web page 1, 73% of the variance of the *Pagerank* variable is accounted for by its linear relationship with *Match*; for data collected concerning web page 2, 80% of the variance of the *Pagerank* variable is accounted for by its linear relationship with *Match*; for data collected concerning web page 3, 47% of the variance of the *Pagerank* variable is accounted for by its linear relationship with *Match*; for data collected concerning web page 4, 63% of the variance of the *Pagerank* variable is accounted for by its linear relationship with *Match*; for data collected concerning web page 5, 100% of the variance of the *Pagerank* variable is accounted for by its linear relationship with *Match*; for data

collected concerning web page 6, 42% of the variance of the *Pagerank* variable is accounted for by its linear relationship with *Match*; for data collected concerning web page 7, 70% of the variance of the *Pagerank* variable is accounted for by its linear relationship with *Match*; for data collected concerning web page 8, 61% of the variance of the *Pagerank* variable is accounted for by its linear relationship with *Match*; for data collected concerning web page 9, 80% of the variance of the *Pagerank* variable is accounted for by its linear relationship with *Match*; for data collected concerning web page 10, 58% of the variance of the *Pagerank* variable is accounted for by its linear relationship with *Match*.

4.2 User Vocabulary and Metadata Vocabulary Compared: Issues Resulting from Differences in Spacing, Spelling, and Other Types of Variations

In this section, "outlier" cases which bucked the trend of the data collected in this study will be highlighted. These cases show that differences between user vocabulary and metadata vocabulary with respect to spacing, spelling, variations of terms, and the addition of terms affected data output. At times such variation between both vocabularies adversely affected the visibility of web pages in search engine results lists.

4.2.1 Web pages 3 and 6

The instances outlined with web pages 3 and 6 highlight important issues that arose as a result of this research study. Data collected concerning web page 3 shows that sometimes exact match did not result in the most effective EWIR. However, these instances were rare. In fact, these were the only two cases in which this occurred out of

all of the data that was collected for all other web pages. It should be noted that, in both cases, the terms employed by the users resulted in retrieved web pages closely related to the desired web page.

As data collected concerning web page 6 shows, the form of terms, singular or plural, employed by users affected match with metadata vocabulary, which, in turn, affected EWIR. As was discussed in the previous section, in the case of web page 6, one user provided the term “fraternities” as a term the user thought would retrieve the Greek life web page. Another user provided the term “sororities” as a term the user thought would retrieve the Greek life web page. In both cases, the desired web page was still visible, but perhaps less visible than otherwise.

4.2.2 Concerning Parts of Words and Abbreviations and Visibility in Search Engine Results Lists

Variations in which the user term was only part of a metadata term resulted in decreased visibility in search engine results lists. An example of a severe case is one in which a subject provided the phrase “Office of undergrad curricula.” This phrase resulted in partial match because the user used the term “undergrad” instead of “undergraduate,” the latter of which was the term used in the metadata vocabulary. Because the user shortened the term, page rank for the desired web page increased from rank 1 to rank 20. This is problematic because it could be argued that undergraduates commonly shorten the term undergraduate with the term undergrad. For example, there

were other instances in which the term “undergrad” was used by subjects in this study (see Appendices 1-10).

In other cases, variation of terms, in the form of abbreviations, did not result in decreased visibility of desired web pages in search engine results lists. (Please note that references to the real name of the study university will be made anonymous by using the word “study”—either with or without full capitalization.) For example, one subject recorded the phrase “STUDY Dept. of Epidemiology” as a phrase he/she thought would yield the study university’s web page for the School of Public Health’s Department of Epidemiology. The abbreviation for department did not adversely affect the visibility of the desired web page in the search engine results list. In fact, using the phrase provided by the user to search for the desired web page resulted in page rank of 1.

4.2.3 Addition of Terms at Times Made EWIR less Effective

Phrases provided by users which varied when compared to metadata vocabulary by the addition of certain terms caused the misfortune of decreased visibility in search engine results lists. For example, one subject wrote the phrase “Fraternity & Sorority (Greek) life at STUDY information.” This phrase resulted in partial match when compared with metadata vocabulary for the study university’s Greek life web page. Out of curiosity, the researcher ran a search with the same phrase minus the last term “information.” Upon completing the modified query, the study university’s Greek life web page was ranked number 1 on the search engine results list. This is problematic because the web page itself is not “Greek life at STUDY.” The web page contains

information about “Greek life at STUDY.” Thus, it seems a quandary if, when users use the term “information” as part of their search query phrase, the term “information” is what would hide a desired web page from their view. In another similar instance, a user provided the phrase “STUDY business school home page” as the phrase he/she believed would most likely return the STUDY business school website with a page rank of 1. Unfortunately, when using the phrase provided by the user, the page rank for the desired web page was 20. Other users (see Appendix 5) provided the same phrase minus the terms “home page,” and, upon providing the phrase “STUDY business school,” the web page was retrieved at a page rank of 1. Again, the actual web page is not the STUDY business school. It contains information about the STUDY business school. It seems unfortunate that by providing a phrase with an additional term that embodies the essence of what the web page actually is, the user would not be able to find that desired web page quickly—if at all.

4.2.4 Concerning Use of Alternate/Generic Terms

Some users provided alternate and perhaps generic terms to express more specific information, and at times this was to no avail. For example, one subject provided the phrase “STUDY daily newspaper” as a phrase he/she thought would yield the enterprise’s school newspaper web page in a search engine results list most visibly and one subject provided the phrase “STUDY campus newspaper” as a phrase he/she thought would yield the enterprise’s school newspaper web page in a search engine results list most visibly. Using these phrases as queries via the enterprise’s search engine yielded the desired web page at a rank of 20. This raises a real issue: what about the newly admitted

undergraduate or transfer student who wants to find the web page for the daily newspaper, but is not aware of the name of the newspaper? How would he/she find it? This is a relevant question to ask, because, based on the results of this study, it seems as though the only way one could find the web page for the newspaper through the enterprise's search engine would be if the user knew what the newspaper was called.

4.2.5 Regarding Variation in Terms of Spacing

Perhaps more alarming, even if one is aware of the name of the newspaper, incorrect spacing among the terms that make up the name of the newspaper can cheat a user out of finding the web page associated with the university newspaper. For example, one subject provided the name of the university newspaper, which is a three word phrase. The only difference is the subject combined the last two words of the phrase. Using the phrase exactly as it was provided by the subject resulted in the desired web page having a page rank of 20. In this case, spacing had a major effect on the visibility of the desired web page. Instances in which users recorded the same terms for the newspaper, but with different spacing, resulted in a page rank of 1 (see Appendix 9). Thus, inaccurate spacing can lead to a significant decrease in visibility of desired web pages in search engine results lists.

4.2.6 Conclusion – Should All of this be Occurring?

Should all of what has been mentioned in this section be happening? Should there be cases in which exact match between user and metadata vocabulary result in less than perfect *EWIR*? Should there be times when using the term "undergrad" instead of

“undergraduate” hides web pages from undergraduates? Should use of words like “information” and “home page” block the visibility of particular web pages from users? Should people who are unaware of the name of their school newspaper be unable to find the web page for their school newspaper? Should mere spacing of words, when everything else is correct, make desired web pages less visible in enterprise search engine results lists? Or should metadata be made more robust by including variations of metadata vocabulary with respect to spelling, spacing, forms of terms, the addition of certain terms that embody the essence of the web page, etc., so that enterprise users will not miss out on finding the web pages they are in search of?

5. Summary and Conclusions

5.1 Summary

In this study, the relationship between user vocabulary and metadata vocabulary was examined as well as what association, if any, existed between that relationship and EWIR. To this end, an enterprise was chosen, an American Research Intensive University, and a sample from a defined set of enterprise users, undergraduates, were recruited. A quasi-experiment was conducted in which twenty undergraduates were asked to view ten web pages, selected from the enterprise's website, and write descriptions for each web page.

The researcher compared the terms written by subjects with metadata terms, which could be viewed via the source code of each of the selected web pages. Instances of exact, partial and no match were recorded by the researcher. Next, the researcher conducted searches using terms supplied by subjects via the enterprise's search engine. The position of the selected web page in the search engine results list was recorded upon completion of each search.

Using SPSS 16.0 the researcher calculated descriptive statistics and bivariate correlation analysis of the data. A statistically significant correlation between the independent variable, match between user vocabulary and metadata vocabulary (*Match*),

and the dependent variable, EWIR, which was operationalized as visibility of web pages via page rank in search engine results lists (*Pagerank*), was found.

For eight out of ten web pages a strong negative correlation between *Match* and *Pagerank* was found. This suggests that, for the most part, when users employ terms that match with terms found in metadata vocabulary for web pages, the page rank of the desired web pages users seek is usually 1 (or close to it) in search engine results lists. This is refreshing because, based on the results of this study, it seems the terms users would employ are liken to the terms metadata creators would use to describe the same web pages. Also, if users used the terms they provided in the quasi-experiment, for the most part, they would find the web pages they were in search of. But only in one instance, in the case of analysis of data related to web page 5, was the negative correlation between *Match* and *Pagerank* "perfect" –which suggests that the transition from users providing terms to finding particular web pages was not always seamless.

5.2 Conclusions

Based on the cases discussed in Chapter 4.2, the researcher suggests that metadata vocabulary could be optimized for the purpose of more effective EWIR by including variations in spacing, spelling, singular and plural variations of terms, and adding words that embody the essence of the web pages. This could increase the chances of user vocabulary matching with metadata vocabulary, and, in turn, increase the effectiveness of metadata for EWIR.

More broadly, the results of this study underscore the reality that words equate to access—at least when considering EWIR. When the words users employ for the purpose

of retrieving web pages become like individual keys, users are unable at times to “open the door” to access the web pages they seek. When variations as minute as spacing place desired web pages out of reach, there is a serious problem. However, metadata vocabulary has the potential to hold “the master key” allowing users the ability to trade in their personal key, their way of describing the web pages they seek and the terms they use to do so, creating a set of circumstances allowing users to retrieve web pages they are looking for in search engine results lists in a highly visible fashion. Thus, providing more robust metadata vocabulary (metadata vocabulary with variation of terms with regard to spelling, spacing, plural and singular form, etc.) could bridge the gap between the terms users provide and the web pages they desire.

The findings in this study suggest that, more often than not, metadata creators are employing terms that users would use with the expectation of retrieving the same web pages for which the metadata was created. However, the terms metadata creators are providing do need work. Metadata vocabulary should be made more robust so that even the slightest variation in user and metadata vocabulary does not result in inaccessibility to desired web pages. Albeit not in the context of a web environment or finding web pages, this is precisely what Furnas et. al. suggested as the appropriate response to what they termed “The Vocabulary Problem in Human-System Communication” over twenty years prior (964, 968).

Works Cited

Babbie, Earl. The Practice of Social Research. 10th ed. Belmont, CA: Thomson/Wadsworth, 2004.

Brasethvik, T. "A semantic modeling approach to metadata." Internet Research: Electronic Networking Applications and Policy. 8.5 (1998): 377–386.

Brooks, Terrence A. "Web Search: how the Web has changed information retrieval." Information Research. 8.3 (2003): 1-14. 27 Jan. 2008 <http://InformationR.net/ir/8-3/paper154.html>.

Dawson, A. & Hamilton, V. "Optimising metadata to make high-value content more accessible to Google users." Journal of Documentation. 62.3 (2005): 307-327.

de Groat, G. "Perspectives on the web and Google: Monika Henziger Director of Research, Google." Journal of Internet Cataloging. 5.1 (2002): 17-28.

Furnas, G. W., Landauer, T.K., Gomez, L.M., & Dumais, S.T. "The Vocabulary Problem in Human-System Communication." Communications of the ACM. 30.11 (1987): 964-971.

Hawking, D. & Zobel, J. "Does topic metadata help with web search?" Journal of the American Society for Information Science and Technology. 58.5 (2007): 613-628.

Henshaw, R. & Valauskas, E.J. "Metadata as a catalyst: experiments with metadata and search engines in the Internet journal, First Monday." Libri. 51.2 (2001): 86-101.

Lynch, C. "When documents deceive: trust and provenance as new factors for

information retrieval in a tangled web." Journal of the American Society for Information Science and Technology. 52.1 (2001): 12-17.

"search engine." *The American Heritage® New Dictionary of Cultural Literacy, Third Edition*. Houghton Mifflin Company, 2005. 04 Apr. 2008. <Dictionary.com http://dictionary.reference.com/browse/search engine>.

Trochim, William M. K., ed. Advances in quasi-experimental design and analysis. San Francisco: Jossey-Bass, 1986.

Turner, T.P. & Brackbill, L. "Rising to the top: evaluating the use of the html meta-tag to improve retrieval of world wide web documents through internet search engines." Library Resources and Technical Services. 42.4 (1998): 258-271.

Zhang, J. & Dimitroff, A. "Internet search engines' response to metadata Dublin Core implementation." Journal of Information Science. 30.4 (2004): 310-320.

----. "The impact of metadata implementation on web page visibility in search engine results (part II)." Information Processing and Management. 41.3 (2005): 691-715.

Appendices 1-11

Appendix 1

Table 1. Web page 1 user vocabulary, *Match* and *Pagerank*

Web page 1 user vocabulary**	*Match*	*Pagerank*
STUDY Dining Services	1	1
STUDY Dining Services	1	1
Dining Services on STUDY campus	0	20
STUDY Dining Services	1	1
food dining	0	20
meal plan, student dining at STUDY	0	3
STUDY dining hall	0	20
STUDY dining hall	0	20
STUDY Dining	1	1
STUDY dining hall	0	20
Dining at STUDY	0	7
Food	-1	20
STUDY dining services	1	1
dining locations	0	20
STUDY Dining services	1	1
STUDY Dining Halls	1	1
STUDY Dining Services	1	1
dining services	1	1
Dining	1	1
STUDY Dining Services	1	1

**In this table web page 1 user vocabulary has been modified such that terms that identified the enterprise, directly or indirectly, were replaced with the term STUDY or alternative descriptions.

Appendix 2

Table 2. Web page 2 user vocabulary, *Match* and *Pagerank*

Web page 2 user vocabulary**	*Match*	*Pagerank*
STUDY Hall	1	1
UNC Lenoir Hall	1	1
STUDY dining hall	1	1
STUDY Dining Hall	1	1
dining hall on campus	0	20
STUDY hall	1	1
STUDY dining hall	1	1
STUDY STUDY dining hall	0	10
STUDY Dining Hall	1	1
STUDY dining hall	1	1
Cafeterias	-1	20
dining hall	0	20
Virtual tour dining hall	1	1
STUDY Hall history	0	20
STUDY STUDY Hall	1	1
STUDY STUDY Dining Hall	0	6
STUDY Dining Hall	1	1
STUDY Hall	1	1
virtual tour	0	20
Lenoir Hall Virtual Tour	1	1

**In this table web page 2 user vocabulary has been modified such that terms that identified the enterprise, directly or indirectly, were replaced with the term STUDY or alternative descriptions.

Appendix 3

Table 3. Web page 3 user vocabulary, *Match* and *Pagerank*

Web page 3 user vocabulary**	*Match*	*Pagerank*
information technology support	0	20
STUDY Information Technology Services	1	1
ITS - Information technology services	1	1
STUDY ITS	1	1
computer problem help	-1	20
computer help	-1	20
STUDY ITS	1	1
STUDY laptops	-1	20
STUDY support	0	20
STUDY help with computers	-1	20
computer services at STUDY	0	1
computer help	-1	20
STUDY ITS	1	1
emergency tech support	-1	20
STUDY ITS help	0	9
computer support	-1	20
STUDY STUDY ITS	0	5
ITS	1	20
Its	1	20
STUDY Information Technology Services	1	1

**In this table web page 3 user vocabulary has been modified such that terms that identified the enterprise, directly or indirectly, were replaced with the term STUDY or alternative descriptions.

Appendix 4

Table 4. Web page 4 user vocabulary, *Match* and *Pagerank*

Web page 4 user vocabulary**	*Match*	*Pagerank*
what is the department of Epidemiology does and is about	0	20
STUDY School of Public Health, Department of Epidemiology	1	1
division of school of health examining and researching health and disease	-1	20
STUDY Dept. of Epidemiology	0	1
news information research	-1	20
school of public health epidemiology	1	1
STUDY epidemiology department	1	1
STUDY department of epidemiology	1	1
STUDY epidemiology	1	1
STUDY department of epidemiology	1	1
Epidemiology	1	1
epidemiology major	0	20
STUDY Department of Epidemiology	0	1
STUDY epidemiology	1	1
STUDY School of Public Health Epidemiology	0	2
Epidemiology	1	1
STUDY Department of Epidemiology	1	1
epidemiology department	1	1
Epidemiology	1	1
STUDY Epidemiology	1	1

**In this table web page 4 user vocabulary has been modified such that terms that identified the enterprise, directly or indirectly, were replaced with the term STUDY or alternative descriptions.

Appendix 5

Table 5. Web page 5 user vocabulary, *Match* and *Pagerank*

Web page 5 user vocabulary**	*Match*	*Pagerank*
global aspect of business	0	20
STUDY Business School	1	1
STUDY business school home page	0	20
STUDY Business School	1	1
business school	1	1
business school	1	1
STUDY business school	1	1
STUDY business school	1	1
STUDY business school	1	1
STUDY business school	1	1
Undergraduate Business	0	20
STUDY business school	1	1
STUDY Business School	1	1
STUDY business major	0	20
STUDY Business School	1	1
STUDY business	1	1
STUDY Business	1	1
business school	1	1
STUDY name of business school	1	1
STUDY name of business school	1	1

**In this table web page 5 user vocabulary has been modified such that terms that identified the enterprise, directly or indirectly, were replaced with the term STUDY or alternative descriptions.

Appendix 6

Table 6. Web page 6 user vocabulary, *Match* and *Pagerank*

Web page 6 user vocabulary**	*Match*	*Pagerank*
History of Greek Life	0	4
STUDY STUDY Fraternity and Sorority	0	1
Fraternity & Sorority (Greek) life at STUDY information	0	20
Fraternities	-1	10
fraternity sorority	1	1
fraternity sorority	1	1
STUDY Greek	0	3
Fraternity STUDY	1	1
STUDY greek	1	1
STUDY frats	0	20
Fraternity/sorority at STUDY	1	1
Fraternities	0	8
Fraternity Life	1	1
STUDY Fraternity and Sorority Life	1	1
STUDY Greek Life	1	1
STUDY Fraternity life	1	1
STUDY Greek	1	1
Greek life	1	1
Sororities	-1	10
STUDY Greek STUDY	1	1

**In this table web page 6 user vocabulary has been modified such that terms that identified the enterprise, directly or indirectly, were replaced with the term STUDY or alternative descriptions.

Appendix 7

Table 7. Web page 7 user vocabulary, *Match* and *Pagerank*

web page 7 user vocabulary**	*Match*	*Pagerank*
Info about residence halls at STUDY	0	20
STUDY Residence Hall Association	1	1
Resident Hall Association	0	20
Residence Hall Association	1	1
Dorming	-1	20
Dorms	-1	20
STUDY dorms	-1	20
STUDY Dorms	-1	20
STUDY housing	0	20
STUDY housing	0	20
Residence Halls for STUDY	0	15
housing info	-1	20
Dorms	-1	20
water conservation at STUDY	0	5
STUDY RHA	1	1
STUDY residence halls	0	16
STUDY Residence Halls	0	16
RHA	1	1
Residence halls	0	20
STUDY RHA	1	1

**In this table web page 7 user vocabulary has been modified such that terms that identified the enterprise, directly or indirectly, were replaced with the term STUDY or alternative descriptions.

Appendix 8

Table 8. Web page 8 user vocabulary, *Match* and *Pagerank*

Web page 8 user vocabulary**	*Match*	*Pagerank*
info about what exactly the undergrad curriculum office does	-1	20
Office of Undergraduate Curricula	1	1
Office of undergrad curricula	0	20
Office of Undergraduate Curricula	1	1
undergrad classes registration course requirements	-1	20
undergrad curricula/curriculum	0	20
STUDY curriculum	-1	20
STUDY Undergraduate Bulletin	0	20
STUDY undergraduate requirements	0	4
STUDY undergrad curriculum	0	20
Undergraduate Curricula	1	1
undergraduate curriculum	0	2
Office of Undergraduate Curricula	1	1
undergraduate curriculum	0	2
STUDY Undergrad curriculum overview	-1	20
STUDY academic advising	-1	20
STUDY undergrad curriculum	-1	20
general curriculum	0	7
undergraduate curricula	1	1
STUDY curriculum	0	20

**In this table web page 8 user vocabulary has been modified such that terms that identified the enterprise, directly or indirectly, were replaced with the term STUDY or alternative descriptions.

Appendix 9

Table 9. Web page 9 user vocabulary, *Match* and *Pagerank*

Web page 9 user vocabulary**	*Match*	*Pagerank*
News pertaining to STUDY	-1	20
Name of STUDY university newspaper	1	1
STUDY daily newspaper	0	20
Name of STUDY university newspaper	1	1
on campus life news paper reporting	-1	20
Name of STUDY university newspaper	1	1
Name of STUDY university newspaper	1	1
Abbreviation of Name of STUDY university Newspaper	1	1
Abbreviation of Name of STUDY university Newspaper	1	1
STUDY campus newspaper	-1	20
STUDY school newspaper	0	20
Name of STUDY university newspaper without spacing	0	20
Abbreviation of Name of STUDY university Newspaper	1	1
STUDY sports news	-1	20
Name of STUDY university newspaper	1	1
STUDY student newspaper	0	20
Name of STUDY university newspaper without spacing	0	20
Abbreviation of Name of STUDY university Newspaper	1	1
Abbreviation of Name of STUDY university Newspaper	1	1
STUDY Newspaper	0	20

**In this table web page 9 user vocabulary has been modified such that terms that identified the enterprise, directly or indirectly, were replaced with the term STUDY or alternative descriptions.

Appendix 10

Table 10. Web page 10 user vocabulary, *Match* and *Pagerank*

Web page 10 user vocabulary**	*Match*	*Pagerank*
Info about what exactly is the environmental science & engineering program	0	20
School of Public Health, Environmental Sciences and Engineering	1	1
Environmental Sciences & Engineering	1	1
STUDY Dept. of Environmental Sciences & Engineering	1	1
public health school environmental engineering graduate programs	0	20
STUDY School of Public Health, environmental science department	1	1
STUDY environmental department	0	2
STUDY environmental science	0	2
STUDY Environmental	0	10
STUDY environment	-1	20
Department of Environmental Sciences	1	1
environmental science major	0	20
Department of Environmental Science and Engineering	1	1
environmental science research	0	6
STUDY School of Public Health enviro sci and engineering	0	20
STUDY public health environmental science department	1	1
STUDY Environmental Sciences & Engineering	1	1
environmental science	0	7
environmental sciences	1	1
Environmental Engineering STUDY	1	1

**In this table web page 10 user vocabulary has been modified such that terms that identified the enterprise, directly or indirectly, were replaced with the term STUDY or alternative descriptions.

Appendix 11

Note: This Instrument appears exactly as it was given to each subject except: 1) the margins of the pages have been increased, which has increased the size of the document, and 2) the urls have been changed to keep the enterprise Anonymous.

Written Query Sheet

Directions:

1) Please type in the url of Web page 1, found below, into the address bar of the Microsoft Internet Explorer 7.0 web browser, which is available from the computer at the computer station.
2) Press enter to view Web page 1.
3) Write the words and/or statements you would use to describe Web page 1 in the space provided below.
4) Repeat steps 1 through 3 for Web pages 2 through 10.

Please spend no more than two minutes per web page to view and write descriptions and please print your responses clearly and legibly. When you are finished, please return the Written Query Sheet to the Principal Investigator (PI).

Web page 1: http://STUDY.edu/url1

Query:

_____

Web page 2: http://STUDY.edu/url2

Query:

_____

Web page 3: http://STUDY.edu/url3

Query:

_____

Web page 4: http://STUDY.edu/url4

Query:

_____

Web page 5: http://STUDY.edu/url5

Query:

_____

Web page 6: http://STUDY.edu/url6

Query:

_____

Web page 7: http://STUDY.edu/url7

Query:

_____

Web page 8: http://STUDY.edu/url8

Query:

_____

Web page 9: http://STUDY.edu/url9

Query:

_____

Web page 10: http://STUDY.edu/url10

Query:

_____

Note: This Written Query Sheet was designed by Devan Ray Donaldson, a Graduate Student in the School of Information and Library Science at the University of North Carolina at Chapel Hill and PI of this experiment. If you have any further questions regarding this experiment, please contact him directly, as he will be supervising the experiment.